Constructing a Watts-Strogatz network from a small-world network with symmetric degree distribution

Mozart B. C. Menezes[1], Seokjin Kim[2]*, Rongbing Huang[3]

1 Operations Management Department, Kedge Business School - Bordeaux, 680 cours de la Libération, 33405 Talence, France, **2** Department of Information Systems and Operations Management, Sawyer Business School, Suffolk University, 120 Tremont Street, Boston, MA 02108, United States of America, **3** School of Administrative Studies, York University, 4700 Keele Street, Toronto, ON, Canada M3J 1P3

* kim@suffolk.edu

Abstract

Though the small-world phenomenon is widespread in many real networks, it is still challenging to replicate a large network at the full scale for further study on its structure and dynamics when sufficient data are not readily available. We propose a method to construct a Watts-Strogatz network using a sample from a small-world network with symmetric degree distribution. Our method yields an estimated degree distribution which fits closely with that of a Watts-Strogatz network and leads into accurate estimates of network metrics such as clustering coefficient and degree of separation. We observe that the accuracy of our method increases as network size increases.

Introduction

Since the term "small world" was coined first by the Milgram's pioneering experiment [1], Watts and Strogatz [2] have proposed the most compelling analytical framework demonstrating the small-world phenomenon prevalent in a range of social, information, technological, and biological networks. A small world consists of many local clusters, but all members are connected with short distance via a few more connected members. These conditions for a small world to emerge are minimal and many real networks have shown small-world properties [3–5]. However, it is not straight-forward to quantify "small-world-ness." A quantitative model measures the equivalence between a network and a unique Watts-Strogatz (WS) model [6]. For a real network with high equivalence, the corresponding WS model can be generated to explore the structure and dynamics.

Even though such equivalence is confirmed, it is not viable to study a large real network when sufficient data on a population are not available. For example, modern social networks are very large in sizes and it takes significant resources and time to collect the data of a network and find key parameters. Web-based experiments accommodating a large number of participants are more difficult to control in some respects than are those conducted in physical laboratories [7]. When such real experiments are impractical, an artificially structured network can

be studied instead [8]. However, a large real network with strong small-world properties cannot be replicated into the corresponding WS model, unless its parameters are estimated from a sample.

A WS model [2] is characterized by n = number of nodes, K = number of neighbors a node has to its right side in the regular lattice before rewiring, and p = "rewiring" probability with which the right end of an arc incident to a node is rewired uniformly randomly to another node. The size of population under study is represented by n which is known a priori or can be estimated [9, 10] in many cases. The sample mean of node degrees is an estimator for K since the total number of arcs remain invariant after rewiring. Among the three parameters, it is the most challenging to estimate p. Motivated by this immediate need, we formulate a method to estimate p leading into an estimated degree distribution which fits closely with that of the corresponding WS model. These three parameters (n, K, p) indeed suffice to characterize a WS network. We observe that, from many generated WS networks under the same values of (n, K, p), variations in network metrics such as clustering coefficient (CC) and degree of separation (DS) (defined as characteristic path length in [2]) are very small.

A direct question from this motivation, then, is how many arcs are incident to node $i \in N$ after rewiring, where N is the set of nodes in a network. We start from deriving the degree distribution of a network represented by the probability $P(\delta_i = m)$ that node i has a degree of m, where P is a probability mass function and δ_i = the degree of node i (or the number of arcs incidents to node i) after rewiring. A previous derivation of $P(\delta_i = m)$ in [11] is based on the assumption that $\delta_i \geq K$ after rewiring, which might not be the case for some WS networks we have generated. Nonetheless, this assumption allows a simpler formulation as a result. We thus propose a new formulation of $P(\delta_i = m)$ which is closer to the exact value in a WS network.

Results and discussion

In the regular lattice of a WS network before rewiring, a node $i \in N$ has degree $2K$ with K arcs incident to its right neighbors and K to its left ones. Let N_i be the set of nodes connected to node i before or during rewiring (whereas δ_i is the degree of node i after rewiring). Then, before rewiring, $|N_i| = 2K$. Note that N_i does not include node i. Node i loses one degree after a sequence of events below takes place to node $j \in N_i$ with the assumption of $|N_i| = 2K$.

1. Arc $\{i, j\}$ is chosen for rewiring with probability p.

2. One end of arc $\{i, j\}$ (attached to node i) is chosen with probability 1/2.

3. The chosen end is rewired, with probability $(n - 1 - 2K)/n$, to a node which is neither node j nor one of the nodes in N_i.

Consequently, the probability that the degree of a node decreases by 1 is $\alpha \equiv \frac{1}{2} p \frac{(n-1-2K)}{n}$. We admit that $|N_i| = 2K$ might not be the case during rewiring and our formulation of $P(\delta_i = m)$ is an approximation. The small world property (high local clustering and short paths) emerges for a small rewiring probability p ranging from 0.001 to 0.1 in Fig 2 in [2]. For a small p, e.g., $p = 0.01$, about 1% of the arcs are rewired. Accordingly, the degree of most nodes would be $N_i = 2K$ during rewiring and this assumption is not significantly limiting. As shown in the examples we have generated, our approximation still results in small errors.

On the other hand, node i gains one degree after the steps below if arc $\{j, k\}$, among the $(n - 2)K$ arcs not incident initially to node i, is detached and an end of arc $\{j, k\}$ is rewired to node i chosen randomly. If $i = j$ or $i = k$, arc $\{j, k\}$ is attached back and not rewired.

4. Arc $\{j, k\}$ is detached for rewiring with probability p.

5. Node i is chosen with probability $1/n$ and an end of arc $\{j, k\}$ is rewired to node i.

Thus, the probability that the degree of a node increases by 1 is $\beta \equiv p/n$.

We rewrite the number of degrees after rewiring as $\delta_i = 2K - X_i + Y_i$, where X_i and Y_i are binomial random variables representing the number of degrees lost and the number of degrees gained at node i, respectively. Then we have

$$P(X_i = x) \quad = \quad \binom{2K}{x} \alpha^x (1-\alpha)^{2K-x}, \quad \text{for } x = 0, 1, ..., 2K; \tag{1}$$

$$P(Y_i = y) \quad = \quad \binom{(n-2)K}{y} \beta^y (1-\beta)^{(n-2)K-y}, \quad \text{for } y = 0, 1, ..., (n-2)K. \tag{2}$$

and

$$P(\delta_i = m) \quad = \quad P(2K - X_i + Y_i = m) \tag{3}$$

$$= \quad \sum_{x=\max\{2K-m,0\}}^{\min\{n-1-m,2K\}} P(Y_i = m - 2K + x \mid X_i = x) \, P(X_i = x). \tag{4}$$

For $m \geq n$ or $m < 0$, $P(\delta_i = m) = 0$.

In Eq (4), we have bounds as $\max\{2K - m, 0\} \leq x \leq \min\{n - 1 - m, 2K\}$, where x is the number of degrees lost. For $0 \leq m \leq 2K$, $2K - m \leq x \leq 2K$. Node i loses at most $2K$ degrees since $|N_i| = 2K$ before rewiring, but cannot lose less than $2K - m$ degrees. Otherwise, $\delta_i = m$ is impossible. For $2K < m \leq n - 1 - 2K$, $0 \leq x \leq 2K$. In this case, node i can lose all $2K$ degrees as long as it can gain from the other $n - 1 - 2K$ nodes and can lose none since $2K < m$. For $n - 1 - 2K < m \leq n - 1$, $0 \leq x \leq n - 1 - m$. Since m is larger than in the previous case, node i loses no more than $n - 1 - m$ degrees and can lose none. In the conditional probability in Eq (4), $Y_i = m - 2K + x$ is immediate from $2K - x + Y_i = m$.

We assume a large n (e.g., $n \gg K$) consistent with a large network which we are mainly interested in. From Eqs (1), (2) and (4), we have

$$P(\delta_i = m) \quad = \quad \sum_{x=\max\{2K-m,0\}}^{\min\{n-1-m,2K\}} \binom{2K}{x} \alpha^x (1-\alpha)^{2K-x} \times$$
$$\binom{(n-2)K}{m-2K+x} \beta^{(m-2K+x)} (1-\beta)^{((n-2)K-(m-2K+x))}. \tag{5}$$

The probability mass function of binomial distribution with probability β in Eq (5) can be approximated by that of a Poisson distribution with rate $\lambda = (n-2)K\beta = (n-2)Kp/n$ for a large n and a small β, which are the cases in small-world networks. From the fact that $\alpha = \frac{1}{2} p \frac{(n-1-2K)}{n} \to \frac{p}{2}$ and $\lambda = \frac{(n-2)}{n} Kp \to Kp$ for a large n, Eq (5) can be written as

$$P(m) = P(\delta_i = m) = \sum_{x=\max\{2K-m,0\}}^{\min\{n-1-m,2K\}} \binom{2K}{x} \left(\frac{p}{2}\right)^x \left(1 - \frac{p}{2}\right)^{2K-x} \times \frac{(Kp)^{m-2K+x}}{(m-2K+x)!} e^{-Kp}. \tag{6}$$

Then the mean estimated from Eq (6) is $\tilde{\mu} = \sum_{m=0}^{n} mP(m)$ and the standard deviation is $\tilde{\sigma} = \sqrt{\sum_{m=0}^{n} (m - \tilde{\mu})^2 P(m)}$. From now on, we use interchangeably use $P(m)$ and $P(\delta_i = m)$.
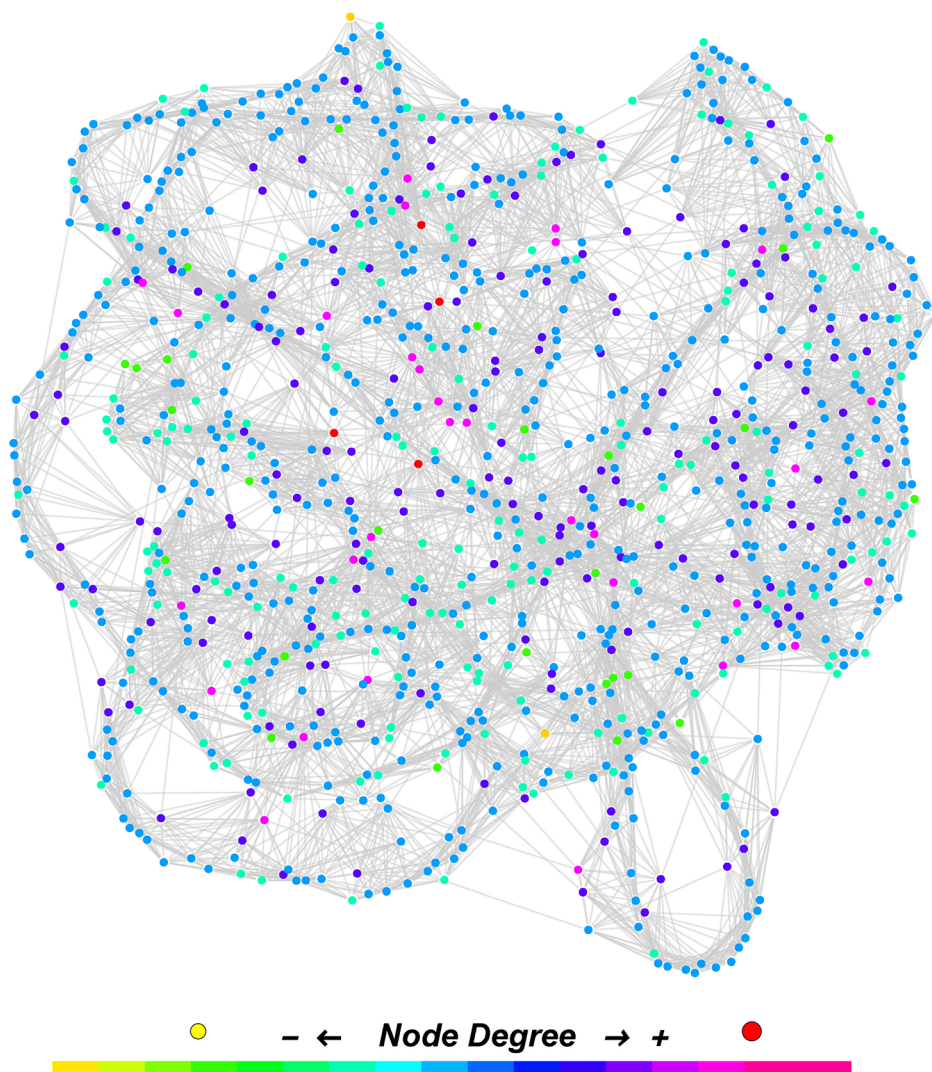
Fig 1. A Watts-Strogatz network. This example with parameters of ($n = 1{,}000$, $K = 8$, $p = 0.04$), generated in Mathematica 10 by Wolfram Research, has $CC = 0.6170$ and $DS = 4.1531$. Nodes are distinguished by different sizes and colors representing their degrees ranging from from 13 to 19.

https://doi.org/10.1371/journal.pone.0179120.g001

Fig 1 includes an actual WS network generated with parameters of ($n = 1{,}000$, $K = 8$, $p = 0.04$). As shown in Fig 2, the degree distribution of the generated WS network is symmetric and closely estimated by Eq (6). Due to the symmetry of degrees in WS networks, our framework is intended for networks with symmetric degree distribution.

The example WS network in Fig 1 is not the only one whose node degrees are close to those estimated by Eq (6). We now demonstrate their statistical fit via 8 tuples of parameters which were set to be $n = 5{,}000$, $10{,}000$, $K = 50$, 75 and $p = 0.01$, 0.05. For each tuple of (n, K, p), 100 WS networks were generated and their node degrees were recorded for two tests performed. First, a chi-square test was performed for each of 800 WS networks between the actual node degrees and estimated values ($nP(\delta_i = m)$) given by Eq (6). None of the 800 tests were significant at the given significance level of $\alpha = 5\%$ and these results corroborate our observation in
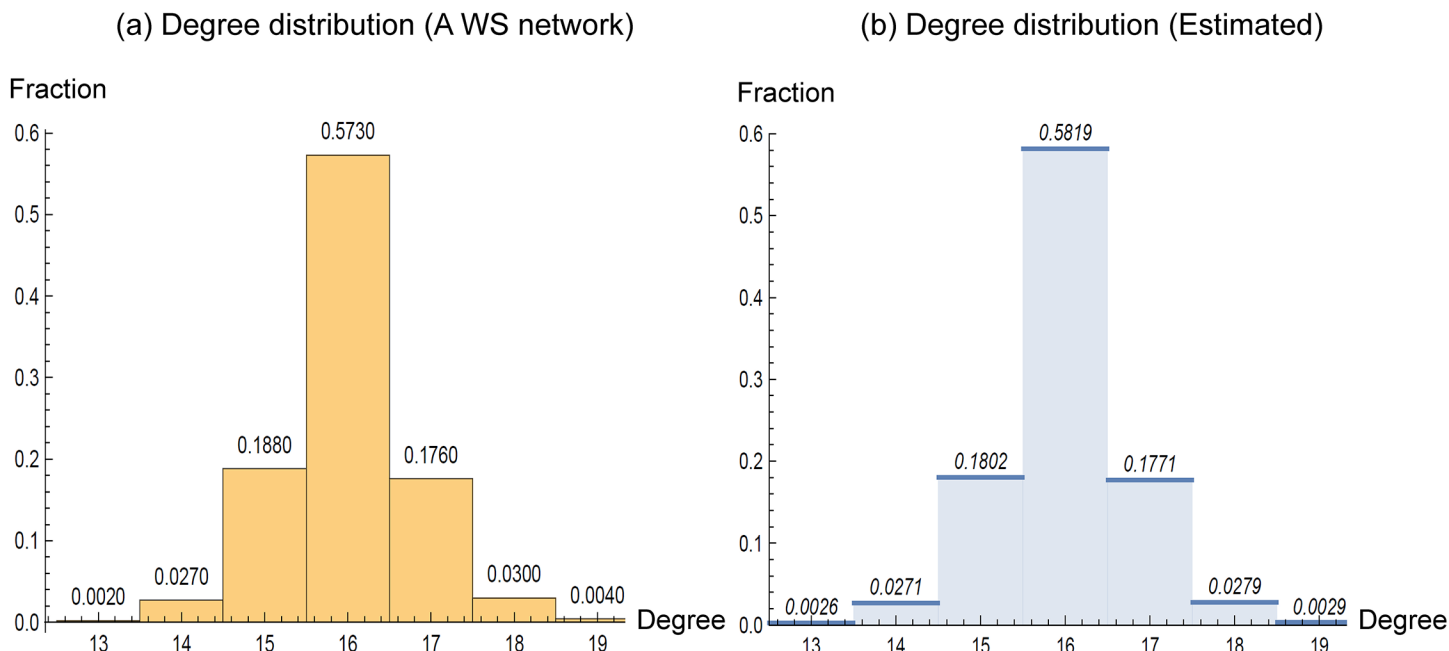
Fig 2. Distributions of degrees in a Watts-Strogatz network and degrees estimated by Eq (6). The WS network was generated with parameters of ($n = 1,000$, $K = 8$, $p = 0.04$).

Fig 2. Second, a t-test was performed on the number of nodes *for each degree m* for 100 WS networks with respect to each tuple of (n, K, p). The 95% confidence interval is $\hat{\mu}_m \pm t_{0.025}(\hat{\sigma}_m/\sqrt{100})$, where $\hat{\mu}_m$ is the sample mean from 100 WS networks and $\hat{\sigma}_m$ is the sample standard deviation. For each tuple and each m, the value of $nP(\delta_i = m)$ estimated by Eq (6) lied within the corresponding confidence interval.

Fig 3 shows the distributions of *CC* and *DS* from 1,000 WS networks randomly generated with parameters of ($n = 1,000$, $K = 8$, $p = 0.04$) kept the same as in Fig 1. The distributions of *CC* and *DS* are symmetric with small ranges ($0.6048 \leq CC \leq 0.6336$ and $4.0278 \leq DS \leq 4.3333$). Thus, given estimates of n, K and p are accurate, the resulting estimates of *CC* and *DS* would also be accurate. It is promising to use our method to estimate K and p from a sample and then evaluate network metrics such as *CC* and *DS* of the corresponding WS network.

Given that n is known or estimated, we propose an algorithm below to find estimates $\hat{K}$ and $\hat{p}$ for their population values of K and p, respectively. Let $S = \{(i, \delta_i); i = 1, \ldots, s\}$ be a set of s individuals sampled from a WS network, where individual i has a degree of δ_i. Since the total number of arcs remains the same after rewiring, an estimate for the sample mean is $\hat{K} = (\sum_{i=1}^{s} \delta_i)/(2s)$ and the sample standard deviation is estimated to be $\hat{\sigma} = \sqrt{(\sum_{i=1}^{s} \delta_i - 2\hat{K})^2/(s - 1)}$. Then we perform a search for $\hat{p}$ until the standard deviation $(\tilde{\sigma}(\hat{p}) = \sqrt{\sum_{m=0}^{n} (m - \tilde{\mu})^2 P(m)}$, where $\tilde{\mu} = \sum_{m=0}^{n} mP(m))$ calculated from Eq (6) gets close enough to $\hat{\sigma}$. Our algorithm is based on the key observation that, as the rewiring probability increases in the WS procedure, the variations of degrees also increase in the resulting network. Thus, given $\tilde{\sigma}(\hat{p})$ from Eq (6), we find $\hat{p}$ in a "reverse" manner.
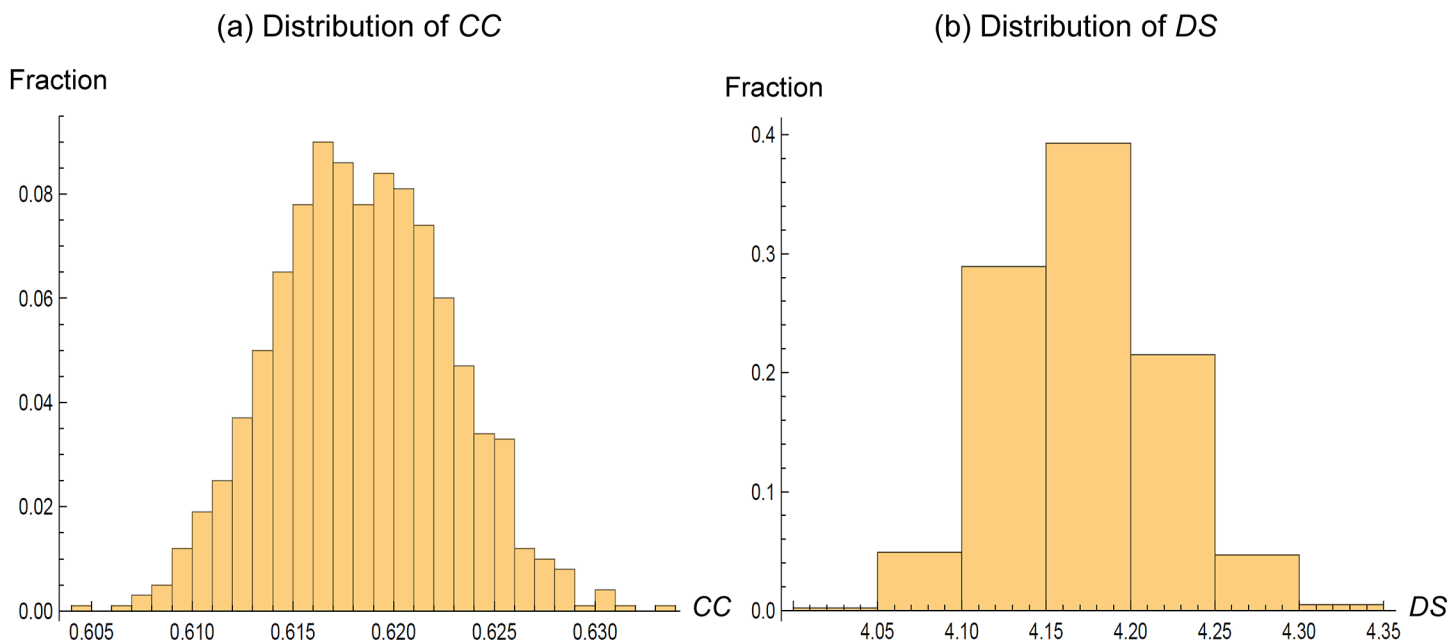
(a) Distribution of *CC*   (b) Distribution of *DS*

Fig 3. Distributions of *CC* and *DS*. 1,000 WS networks were randomly generated with parameters of ($n = 1000$, $K = 8$, $p = 0.04$).

Algorithm 1

Input: n, $(\delta_i, i = 1, \ldots, s)$
Define ϵ with a very small value (e.g., 0.00001).
Calculate $\hat{K} = \sum_{i=1}^{s} \delta_i/(2s)$ and $\hat{\sigma} = \sqrt{\sum_{i=1}^{s} (\delta_i - 2\hat{K})^2/(s-1)}$.
Let $p_l = 0$, $p_u = 1$ and $\hat{p} = (p_l + p_u)/2$.
Given $\hat{p}$, use Eq (6) to calculate $\tilde{\mu} = \sum_{m=0}^{n} mP(m)$ and $\tilde{\sigma}(\hat{p}) = \sqrt{\sum_{m=0}^{n} (m - \tilde{\mu})^2 P(m)}$.
while $\tilde{\sigma}(\hat{p}) \notin [\hat{\sigma} - \epsilon, \hat{\sigma} + \epsilon]$ **do**
   If $\tilde{\sigma}(\hat{p}) > \hat{\sigma}$, let $p_u = \hat{p}$. Else, let $p_l = \hat{p}$.
   Let $\hat{p} = (p_l + p_u)/2$ and use Eq (6) to calculate $\tilde{\mu}$ and $\tilde{\sigma}(\hat{p})$.
end while
Construct a WS network with parameters of $(n, \hat{K}, \hat{p})$, and calculate *CC* and *DS*.
Output : $(n, \hat{K}, \hat{p}), CC, DS$

Fig 4 shows results from application of Algorithm 1 to 200 WS networks (100 networks for $n = 10,000$ and 100 networks for $n = 20,000$) with $K = 80$ and a randomly chosen $p \in [0.005, 0.05]$. From each network, $s = 100$ nodes were randomly sampled along with their degrees to calculate $\hat{p}$ from Algorithm 1. The closer the labels are to the diagonal, the more accurate the estimated values match the actual ones. In Fig 4, we normalized all estimated values between 0 and 1 to plot them together. Closer matches between the estimated and actual values are observed for larger networks with $n = 20,000$. The variations around the diagonal seem to be consistent with those in Fig 3, but exacerbated slightly due to the use of $\hat{p}$ as an estimate of p.

In Fig 5, we measure the accuracy of $\hat{p}$ calculated by Algorithm 1. For each combination of $n = 10,000, 20,000$, $K = 80$ and s/n (percentage of nodes sampled) = 1%, 3%, 5%, we generated 30 WS networks with a randomly chosen $p \in [0.05, 0.2]$ and calculated $\hat{p}$ from Algorithm 1. Then, for each combination, we calculated the sample mean of 30 ratios $(\hat{p}/p)$. The ratio of 1 represents an exact match between $\hat{p}$ and p. As sample sizes increase from 1% to 5% of nodes sampled, mean ratios of $\hat{p}/p$ approach to 1. Also, as in Fig 4, higher accuracy is observed for
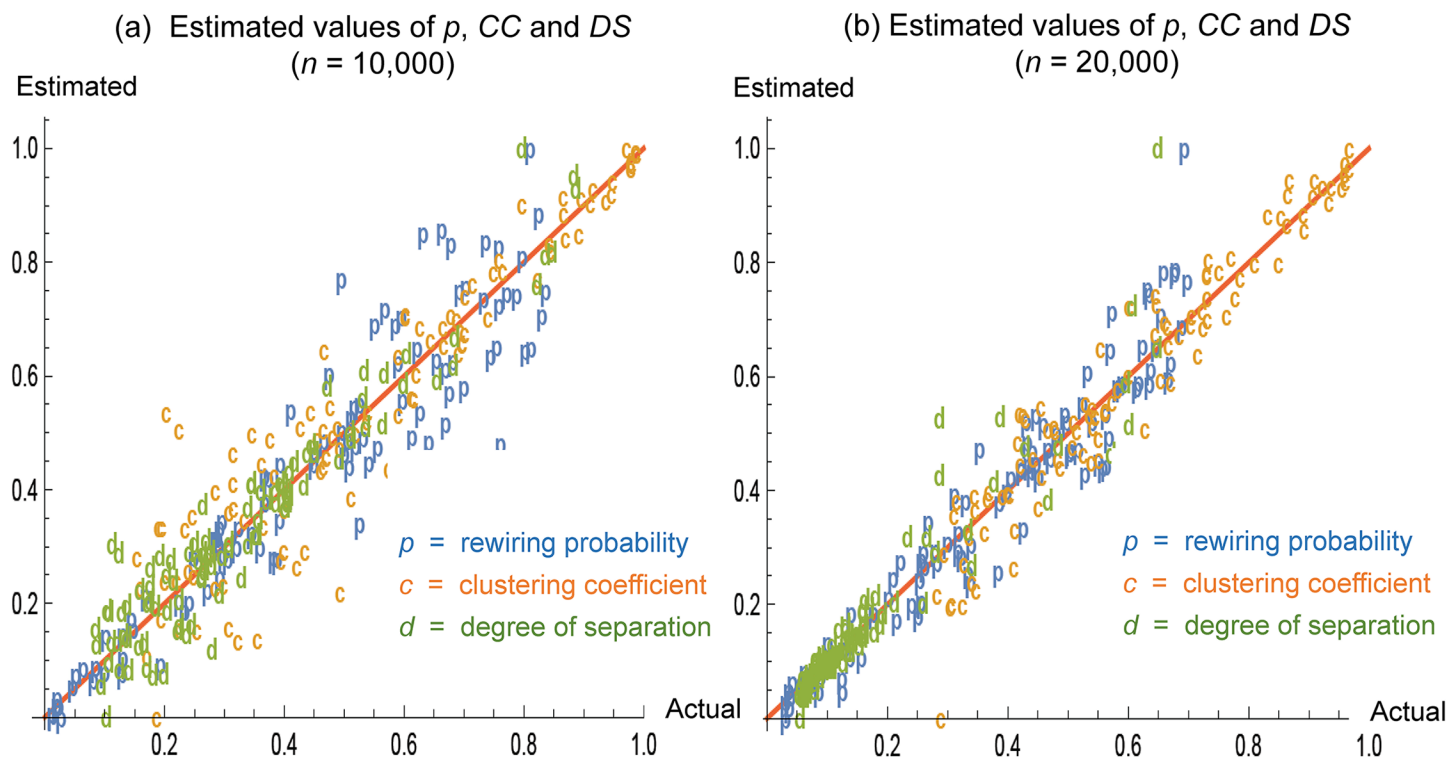
Fig 4. Estimated values of _p_, _CC_ and _DS_. For each case of (a) $n = 10,000$ and (b) $n = 20,000$, 100 WS networks were generated with $K = 80$ and a randomly chosen $p \in [0.005, 0.05]$. From each network, $s = 100$ nodes were randomly sampled along with their degrees to calculate $\hat{p}$ from Algorithm 1. All estimated values were normalized between 0 and 1.

larger networks with $n = 20,000$. For each percentage of nodes sampled ($s/n = 1\%, 3\%, 5\%$), the confidence interval for $n = 20,000$ is narrower than that for $n = 10,000$ while both of the confidence intervals overlap with $\hat{p}/p = 1$.

Table 1 summarizes the sample means and 95% confidence intervals (in a format of sample mean ± margin of error) in Fig 5. A margin of error is calculated as $t_{0.025}(\sigma_r/\sqrt{30})$, where σ_r is the sample standard deviation of 30 ratios. Again, For each percentage of nodes sampled ($s/n = 1\%, 3\%, 5\%$), the margins of error for $n = 20,000$ are smaller resulting in narrower confidence intervals. Thus, our method adds more accuracy for larger networks (e.g., large-scale social networks).

Conclusion

We have presented a method to construct a Watts-Strogatz network using a sample from a small-world network with symmetric degree distribution. Our method yields an estimated degree distribution which fits closely with that of a WS network and allows to characterize the population with accurate estimates of network metrics such as clustering coefficient and degree of separation. This is particularly useful when sufficient information on the population is not available due to limited resources and time. As observed, our method is more accurate for larger networks.

An obvious limitation of our method is the symmetry of degree distribution and we admit that many real networks have skewed degree distributions. Applications of our method are
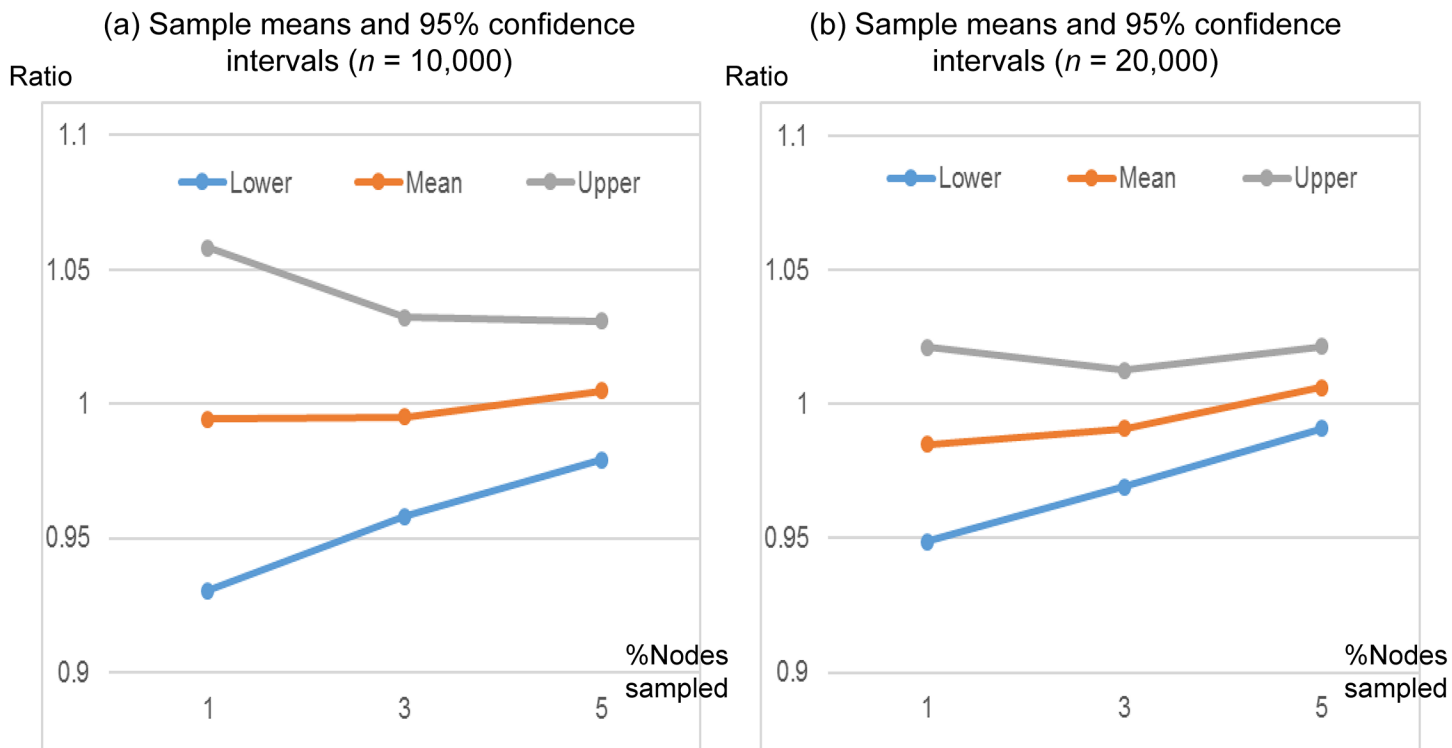
(a) Sample means and 95% confidence intervals (n = 10,000)

(b) Sample means and 95% confidence intervals (n = 20,000)

Fig 5. Sample means and 95% confidence intervals of 30 ratios of $\hat{p}/p$. For each combination of n = 10,000, 20,000, K = 80 and s/n = 1%, 3%, 5%, 30 WS networks were generated with a randomly chosen $p \in [0.05, 0.2]$. For each network, the value of $\hat{p}$ was calculated from Algorithm 1.

https://doi.org/10.1371/journal.pone.0179120.g005

Table 1. Sample means and 95% confidence intervals of 30 ratios of $\hat{p}/p$.

n	s/n = 1%	s/n = 3%	s/n = 5%
10,000	0.9943 ± 0.0638	0.9952 ± 0.0371	1.0050 ± 0.0258
20,000	0.9850 ± 0.0362	0.9908 ± 0.0218	1.0061 ± 0.0152

Table notes 95% confidence intervals are shown in a format of sample mean ± margin of error. For each combination of n = 10,000, 20,000, K = 80 and s/n = 1%, 3%, 5%, 30 WS networks were generated with a randomly chosen $p \in [0.05, 0.2]$. For each network, the value of $\hat{p}$ was calculated from Algorithm 1.

https://doi.org/10.1371/journal.pone.0179120.t001

also limited to networks revealing strong small-world properties which can be well represented by a WS model, since our method is formulated based on the WS rewiring procedure. For a real network either with non-symmetric degree distribution or with weak small-world properties, we still hope that our method serves as a building block for potential revisions or extensions.

Replicating a large network from a sample allows further experiments on a generated network for more insights on its structure and dynamics. This would be feasible if some fundamental properties (e.g., small-world properties) of the network are identified and formulated in an analytical model (e.g., the WS model). Then, key parameters of the model can be estimated and a full-scale network can be constructed.

Author Contributions

Conceptualization: MM SK RH.

Data curation: MM SK.

Formal analysis: MM SK RH.

Investigation: MM SK.

Methodology: MM SK RH.

Project administration: SK MM RH.

Resources: MM SK RH.

Software: MM SK.

Supervision: SK MM RH.

Validation: SK MM RH.

Visualization: MM SK RH.

Writing – original draft: SK MM RH.

Writing – review & editing: SK MM RH.

References

1. Milgram S. The small-world problem. Psychol Today. 1967 1(1):61–67.

2. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1988 Jun; 393:440–442. https://doi.org/10.1038/30918

3. Amaral LAN, Scala A, Barthélémy M, Stanley HE. Classes of small-world networks. Proc Natl Acad Sci U S A. 2000 97(21):11149–11152. https://doi.org/10.1073/pnas.200327197 PMID: 11005838

4. Newman MEJ. The structure and function of complex networks. SIAM Rev Soc Ind Appl Math. 2003 45 (2):167–256.

5. Schnettler S. A structured overview of 50 years of small-world research. Soc Networks. 2009 31:165–178. https://doi.org/10.1016/j.socnet.2008.12.004

6. Humphries MD, Gurney K. Network 'Small-World-Ness': A quantitative method for determining canonical network equivalence. PLoS One. 2008 3(4):e0002051. https://doi.org/10.1371/journal.pone.0002051 PMID: 18446219

7. Salganik MJ, Dodds PS, Watts DJ. Experimental study of inequality and unpreditability in an artificial cultural market. Science. 2006 311:854–856. https://doi.org/10.1126/science.1121066 PMID: 16469928

8. Centola D. The spread of behavior in an online social network experiment. Science. 2010 329:1194–1197. https://doi.org/10.1126/science.1185231 PMID: 20813952

9. Killworth PD, Johnsen EC, Russell Bernard H, Shelley GA, McCarty C. Estimating the size of personal networks. Soc Networks. 1990 23:289–312. https://doi.org/10.1016/0378-8733(90)90012-X

10. Maltiel R, Raftery AE, McCormick TH, Baraff AJ. Estimating population size using the network scale up method. Ann Appl Stat. 2015 9(3):1247–1277. https://doi.org/10.1214/15-AOAS827 PMID: 26949438

11. Barrat A, Weigt M. On the properties of small-work network models. Eur Phys J B. 2000 13:547–560. https://doi.org/10.1007/s100510050067